

Subset-Based Ant Colony Optimisation for the Discovery of Gene-Gene Interactions in Genome Wide Association Studies

Emmanuel Sapin
University of Exeter, College of
Engineering,
Mathematics and Physical
Sciences, Harrison Building
Exeter, England
e.sapin@exeter.ac.uk

Ed Keedwell
University of Exeter, College of
Engineering,
Mathematics and Physical
Sciences, Harrison Building
Exeter, England
e.c.keedwell@exeter.ac.uk

Tim Frayling
Genetics of Complex Traits,
Institute of Biomedical and
Clinical Science, Peninsula
Medical School,
Magdalen Road, Exeter,
England
tim.frayling@pms.ac.uk

ABSTRACT

In this paper an ant colony optimisation approach for the discovery of gene-gene interactions in genome-wide association study (GWAS) data is proposed. The subset-based approach includes a novel encoding mechanism and tournament selection to analyse full scale GWAS data consisting of hundreds of thousands of variables to discover associations between combinations of small DNA changes and Type II diabetes. The method is tested on a large established database from the Wellcome Trust Case Control Consortium and is shown to discover combinations that are statistically significant and biologically relevant within reasonable computational time.

Categories and Subject Descriptors

J.3 [Life and Medical Sciences]: Biology and Genetics; G.1.6 [Numerical Analysis]: Optimization—*Stochastic programming, Global optimization*

Keywords

Ant Colony Optimisation, Genome Wide Association, Type II Diabetes, Single-Nucleotide Polymorphisms

1. INTRODUCTION

The entire human genome has been known for less than a decade, and yet next-generation sequencing methods mean that sequencing genomes has never been cheaper, and are about to fall below the \$1,000 barrier. Researchers therefore are now able to sequence many hundreds or thousands of individuals, allowing an unprecedented view of disease and variation in traits across populations. These studies,

known as genome-wide association studies (GWASs), investigate the relationship between small changes in the genome called Single Nucleotide Polymorphisms (SNPs) and the propensity to develop certain diseases such as Type I and Type II diabetes. These studies offer the potential to illuminate the genetic causes of disease and provide an opportunity for early treatment and planning for patients leading to profound social and economic benefits.

From a computational perspective GWASs present a significant challenge as the number of variables per individual is large (approximately 500,000 in this study) per individual and in the majority of studies these are recorded for thousands of individuals. Any computational approaches used to analyse these data therefore must be scalable in the face of these large-scale data. This paper presents a subset-based tournament ant colony approach to the analysis of full-scale GWAS data with the aim to find combinations of SNPs (gene-gene interactions) that have association with Type II diabetes across of a population of 5,003 individuals.

1.1 Previous Research

Many examples of GWAS data analysis exist in the literature that demonstrate the association between a single SNP and disease. There are known single associations for Type II diabetes and traits such as height for instance. However, there is a considerable amount of missing heritability, for example only about 10% of variation in height can be explained by traditional GWAS. This missing heritability could be due to rare variants, or to the combination of SNPs (gene-gene interactions) which have rarely been explored in the past. Standard GWAS analyses are carried out through full enumeration and with modern hardware, the association of hundreds of thousands SNPs with a disease can be determined within reasonable computational time. However, when combinations (pairs, triplets and higher) are considered, the computational load becomes highly burdensome or completely intractable. This has led to the adoption of metaheuristics [13] for this purpose and previous research has shown ant colony optimisation (ACO) to be a promising technique [7] [1].

ACO has also been shown to deliver excellent results on discrete combinatorial test problems [3] and has been widely applied to real-world problems ranging from water distribution system optimisation [19, 17] to bioinformatics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

GECCO'13, July 6–10, 2013, Amsterdam, The Netherlands.
Copyright 2013 ACM 978-1-4503-1963-8/13/07 ...\$15.00.

It is worth noting that here, a metaheuristic is successfully applied to a full set of GWAS data, comprising many SNPs and individuals without expert knowledge. Whereas previously ACO has only been applied to a single chromosome of real data [1] or with expert knowledge and test data [7].

1.2 Type II Diabetes

Type II diabetes (T2D), characterized by insulin resistance, affects more than 170 million people worldwide [12] and the heritability of this disease has been shown widely in previous studies [16]. Zeggini et al. [20] provide insight into the genetic architecture of T2D detecting replicated diabetes association signals. Gloyn et al. [6] evaluated the role of genetic variants to confirm if they increase risk of T2D and Grant et al. [14] identified a common genetic variant in Type II diabetes mellitus and it does not appear to be in a coding region. However identifying genetic risk in T2D has met with only limited success [8] and therefore remains a formidable challenge [10].

1.3 Data

This research uses the database of The Wellcome Trust Case Control Consortium (WTCCC) which is a collection of GWAS studies relating to a variety of diseases including Type II Diabetes [2]. In this database, a total of 5003 human genomes are provided, with ~500,000 SNPs recorded for each individual in the database. Each SNP represents a small change in the genome and consists of two alleles (two letters among A, C, G, T). Due to the diploid nature of human genomes, there are three possible genotypes for each SNP (e.g. CC, GG, CG).

2. METHOD

2.1 Memory Management

The database on which the research is based is composed of a sample of the genome of 1,999 individuals with Type II diabetes (Cases) and 3,004 control individuals. Each sample of genome is composed of 490,294 single-nucleotide polymorphisms (SNPs) that can have the values unknown, common homozygous e.g. CC, heterozygous e.g. CG, rare homozygous e.g. GG. The sample of the genome of the 5,003 individuals were stored on a hard drive in files. It required 95 minutes to open and to read these files using a machine with a 1TB 7200RPM hard-drive and an Intel Core i7-2600 CPU @3.40GHz processor.

An ACO run may require over a million fitness function evaluations and so even a small improvement in complexity of this function will have a large impact on performance. A function that requires the searching of a database on disk will lead to run times orders of magnitude longer than one in which the data are stored in memory. However, with more than 2 billion elements to represent (5,003 individuals \times 490,294 SNPs), each SNP cannot be represented by more than 1-2 bytes in memory for normal desktop memory sizes (the above machine has 8GB RAM). The representation below implements a lossless compression of the data that enables it to be kept in memory when considering whole genome analysis.

In order to reduce the time to access to the database, two files 'diabetes' and 'controls' were generated. In these two files, the four values 0, 1, 2 and 3 for SNPs encode respectively unknown, common homozygous, heterozygous

and rare homozygous, such that four SNPs of an individual are encoded in one byte ($4^4 = 256$). For example, if *snp1* is heterozygous, *snp2* is unknown, *snp3* is common homozygous and *snp4* is rare homozygous, then the encoding is $147 (64 \times 2 \text{ (Heterozygous)}) + 16 \times 0 \text{ (Unknown)} + 4 \times 1 \text{ (Common homozygous)} + 1 \times 3 \text{ (Rare homozygous)}$.

Using this encoding only 1 minute 13 seconds is required to open and to read the files 'diabetes' and 'controls' using the same machine specification.

Using this representation the size of the database is $490,294 (\text{number of SNPs}) \times 5,003 (\text{number of individuals}) / 4 (\text{number of SNPs per byte}) = 0.585 \text{ Gigabytes}$ therefore it is possible to store the database entirely in RAM, significantly reducing loading times and greatly reducing the run time of the ACO method.

2.2 Exclusion of SNPs

GWAS data have to undergo a series of quality control tests before it can be used. Therefore there are some exclusion criteria that must be applied to ensure the quality of the remaining data. The SNPs that were kept are those that met these four conditions in the 3,004 control samples. Readers are asked to refer to the GWAS literature for more information on these criteria [5].

- Hardy-Weinberg equilibrium Exact Test $> 10^{-4}$
- Minor allele frequency $> 1\%$ for these 3,004 individuals
- Studywise missing data proportion $< 5\%$ for these 3,004 individuals
- Studywise minor allele frequency $> 5\%$ for these 3,004 individuals or studywise missing data proportion $< 1\%$ for these 3,004 individuals

and meeting these two conditions in the 1,999 samples of genomes of individuals with Type II diabetes:

- Hardy-Weinberg equilibrium Exact Test $> 10^{-4}$
- Minor allele frequency $> 1\%$ for these 1,999 individuals.

The resulting filtered dataset contains 405,139 SNPs.

2.3 Tournament ACO Algorithm

As described above, this paper is concerned with the combinations of two SNPs that can discriminate Type II diabetes (T2D). The type of combination that is considered here is the following: An individual is positive if and only if a first SNP (*snp1*) takes a specific value (*value1*) and a second SNP (*snp2*) takes a specific value (*value2*). Ant Colony Optimisation (ACO) is used to search for combinations of SNPs that can discriminate type II diabetes (T2D) from controls in a fashion similar to binary classification.

This search is described as a subset problem as there is no concept of order between the two SNPs. Early experiments with ACO algorithms concentrated on ordering problems such as the Quadratic Assignment Problem or the Traveling Salesman Problem. With the latter being used as example to illustrate the basic concepts of the original Ant System [11].

However, Leguizamon and Michalewicz showed that ACO can be applied to subset problems [9]. Ordering problems

and subset problems differ in the way the pheromone trail is laid on the construction graph. For ordering problems, the connections between elements of the problem receive the pheromone whereas the components themselves receive pheromone in subset problems. The ACO algorithm uses the representation described by Leguizamón and Michalewicz as the ordering of SNPs is not important in the problem formulation.

2.4 Algorithm

Ant Colony Optimisation (ACO) is used to search for combinations of SNPs that can discriminate the T2D and control populations. A value is associated to every SNP that represents how good the combination of SNPs is at discriminating between these two classes. This value $P(n)$ is called the amount of pheromone of the SNP n . The amount of pheromones P is used to guide a tournament selection process to select SNPs for new combinations.

The algorithm can be described as follow:

```

1 Initialise pheromone on each SNP
2 Repeat
3   For all the nbant ants:
4     Select two SNPs via tournament
      selection (see section below)
5     Calculate the fitness of the combination
6   End
7   Updated pheromone of the two SNPs with the
  best fitness
8   For all SNPs: apply evaporation rate e
9 End

```

Where:

- *nbant*: numbers of ants of the algorithm;
- *e*: evaporation rate;

2.4.1 Initialisation of the Pheromone Trail

The pheromone trails were initialised to the highest allelic odds ratio of each individual SNP, although in additional experiments the uniform initialisation of trails to 1 was found to produce similar results.

2.4.2 Tournament Selection

Selecting SNPs for new combinations based on their pheromone value is a key point of the ACO algorithm. Standard ant colony optimisation algorithms use a roulette wheel type structure for this purpose but for 405,139 SNPs, the implementation of a roulette wheel will slow down the algorithm and a tournament selection was proven to be better for high dimension problems [4]. In the tournament-based approach a number of *nbt* SNPs are randomly selected from the possible set to form a tournament and the SNP with the highest amount of pheromone among them is selected as part of this solution. This is the standard method for tournament selection as used in evolutionary algorithms and the tournament size t can be varied to alter the speed of convergence of the algorithm.

2.4.3 Fitness Function

The value of fitness function depends on the two chosen SNPs *snp1* and *snp2* and the two values *value1* and *value2*. For each control and case the following test is performed.

$snp1 = value1$ and $snp2 = value2$

The numbers of positive and negative controls C_p and C_n and the number of positive and negative cases D_p and D_n are determined. The expected values $E.C_p$, $E.C_n$, $E.D_p$ and $E.D_n$ for C_p , C_n , D_p and D_n are calculated as shown in table 1.

The chi square for *value1* and *value2* is given by the formula:

$$Chi^2_{value1,value2} = \frac{(E.D_p - D_p)^2}{E.D_p} + \frac{(E.D_n - D_n)^2}{E.D_n} + \frac{(E.C_p - C_p)^2}{E.C_p} + \frac{(E.C_n - C_n)^2}{E.C_n}.$$

There are three possible values (Common Homozygous, Heterozygous, Rare Homozygous) for *value1* and again three possible values for *value2*. Therefore there are 9 (3×3) different chi square values and the largest of these is selected as the value f of the fitness function of the combination of the two SNPs *snp1* and *snp2*.

$f = \max\{Chi^2_{value1,value2}\}$ such that $(value1,value2) \in \{\text{Common Homozygous, Heterozygous, Rare Homozygous}\}^2$

From the chi square the p-value (probability of achieving this result through chance) of the association can be calculated.

2.4.4 Updating pheromone

The amounts of pheromone of the two SNPs containing in the combination with the highest fitness f among the *nbant* combinations (for each ant a combination of two SNPs is chosen) are updated. For the two amounts of pheromones the following are applied:

$$P(snp1) = P(snp1) + f$$

$$P(snp2) = P(snp2) + f$$

2.4.5 Evaporation

For all SNPs, the following equation with an evaporation rate of 1% is applied:

$$P(snp) = P(snp) \times 0.99$$

3. EXPERIMENTATION

3.1 Monte Carlo Approach

A key aspect of the analysis of GWAS data is that any discovered associations should be checked to determine the likelihood that they could have been generated by chance. The statistics presented in the tables in later sections provide statistical measures of this by calculating odds ratios, chi-squared and p-values but to gain a more in-depth understanding of the performance of the ACO, and the probability of discovering these results by chance, one million random pairs of SNPs were generated and tested using the above fitness function. The average p-value of these pairs is 3.5×10^{-7} , the best is 1.9×10^{-13} . Figure 1 shows the distribution of the p-value of the one million combinations of SNPs.

3.2 Parameters

The key parameters of the ACO algorithm are:

- *nbant*: Numbers of ants of the algorithm
- *nbt*: Number of SNPs in the tournament of the selection process.

An investigation into these parameters was conducted to determine the effect of changing the population size and

	Cases	Controls	Total
Positive	D_p	C_p	$D_p + C_p$
Negative	D_n	C_n	$D_n + C_n$
Total	$D_p + D_n$	$C_p + C_n$	$D_p + D_n + C_p + C_n$
Expected Positive	$E.D_p = \frac{(D_p+D_n) \times (D_p+C_p)}{D_p+D_n+C_p+C_n}$	$E.C_p = \frac{(C_p+C_n) \times (D_p+C_p)}{D_p+D_n+C_p+C_n}$	$D_p + C_p$
Expected Negative	$E.D_n = \frac{(D_p+D_n) \times (D_n+C_n)}{D_p+D_n+C_p+C_n}$	$E.C_n = \frac{(C_p+C_n) \times (D_n+C_n)}{D_p+D_n+C_p+C_n}$	$D_n + C_n$
Total	$D_p + D_n$	$C_p + C_n$	$D_p + D_n + C_p + C_n$

Table 1: Calculation of the expected values $E.C_p$, $E.C_n$, $E.D_p$ and $E.D_n$ for C_p , C_n , D_p and D_n .

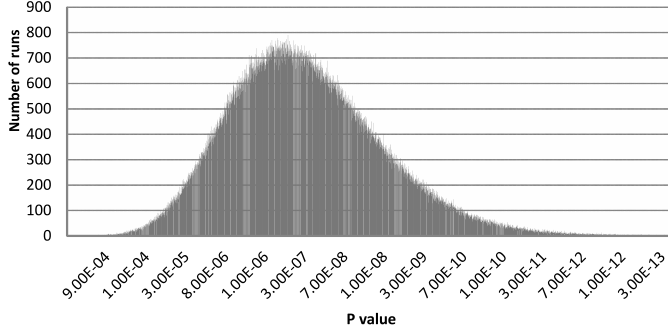


Figure 1: Distribution of the best p-value of 1 million combinations of two SNPs randomly selected.

tournament size on the execution of the algorithm. Inspired by [4], the algorithm was run with the values 100, 200 and 500 for the number of ants and the values 20, 50 and 100 for the tournament size. 10 algorithm runs were conducted for each combination of values of these two parameters. For each run, the best chi squared value against time are shown in figure 2. The results are as expected in that the larger tournaments lead to earlier convergence than smaller ones, as larger tournaments bias the search towards SNPs that have received pheromone early in the search process. However, due to the large-scale nature of the problem, a tournament size of 100 still only represents 0.025% of the variables in the dataset.

The number of ants in the population does not significantly affect the result and so 100 was chosen. For a tournament size of 20, the best chi square is lower (worse) than for 50 and 100 items, therefore 50 was chosen for the number of items because the choice of 100 items would slow down the execution of the algorithm.

4. RESULTS

The best results of 10 algorithm runs over 10000 generations with 100 ants and 50 items in the tournament selection are saved into files. With these parameters, a generation of the algorithm lasts on average 1.28 seconds so a run lasts an average of 3 hours and 32 minutes using a machine with a processor and a hard drive as outlined in the method section.

The ACO algorithm found good results for combinations of two SNPs. The first combination presented here has not previously been known to discriminate T2D where the second is known to discriminate T2D. The last subsection presents a sample of the best combinations that were discovered.

Rs number	position	Chromosome		Allelic OR
7031174	36651528	9		1.178
		CC	CA	AA
Diabetics	1397	443	15	
Controls	2266	557	37	

Table 2: SNP of rs number 7031174.

Rs number	position	Chromosome		Allelic OR
7045471	36603003	9		0.821
		AA	AG	GG
Diabetics	1647	321	17	
Controls	2388	565	37	

Table 3: SNP of rs number 7045471.

	Cases	Controls
Positive	119	11
Negative	1732	2865
Expected Positive	50.91	79.09
Expected Negative	1800	2796.91
OR	17.89	
95% CI	9.6 to 33	
χ^2	154	
P-value	2×10^{-35}	

Table 4: Result of the association rs7031174 = Het. and rs7045471 = Com. Hom..

4.1 Combination of new SNPs

Tables 2 and 3 describe the SNPs in isolation and provide an idea of how discriminatory they are with respect to the cases and controls. The allelic odds ratio (Allelic OR) provides a measure of discrimination between the two classes and values significantly different to 1.0 indicate a discriminatory capability.

Table 4 shows the impact of the combination of the SNPs and the greatly increased odds ratio (now 17.89) indicating that the SNPs are far more discriminatory in combination than they are in isolation, a phenomenon known as epistasis. The table also shows the expected and actual values for each of the categories where *positive* indicates those individuals where the condition is true and *negative* indicates those individuals for whom the condition is false. As can be seen more control individuals are present in the positive category than would be expected given the distribution of cases and controls in the dataset and so this combination of SNP values (rs7031174 = CA and rs7045471 = AA) can be thought of as being negatively associated with T2D. The statistical evidence suggests that this is a strong result with a p-value (the probability of achieving this result through chance) of 2×10^{-35} .

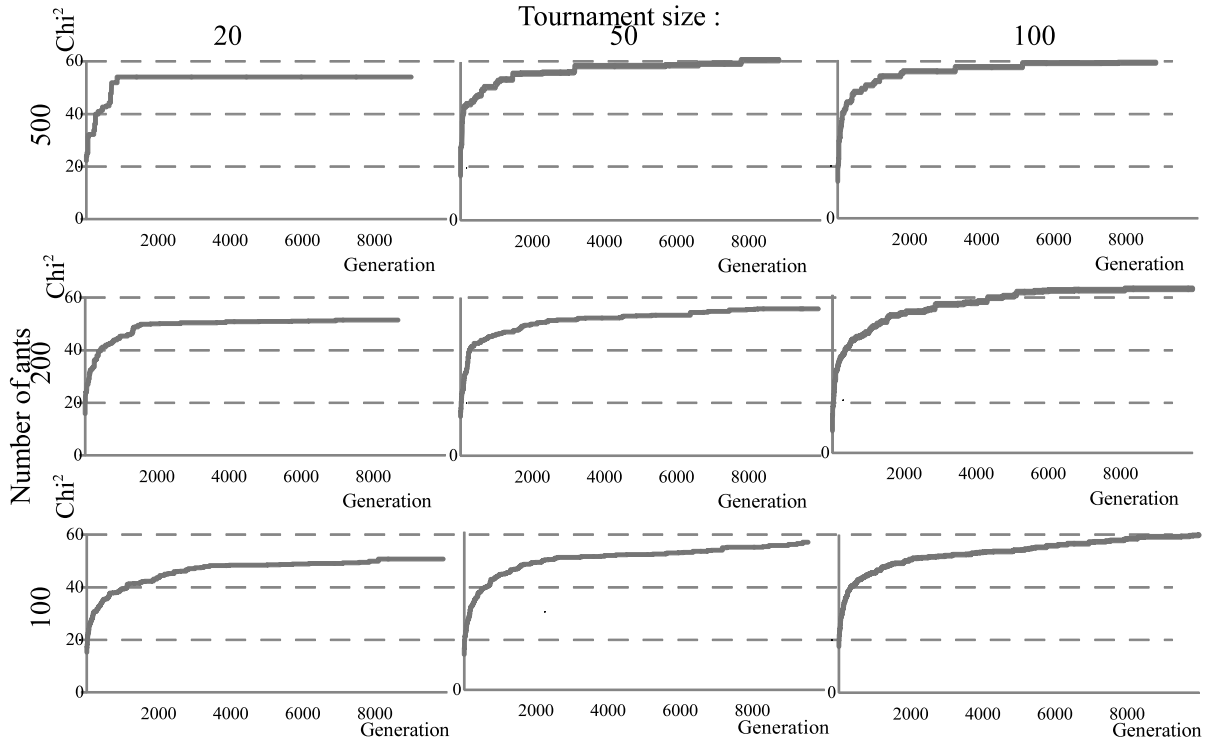


Figure 2: Evolution of the average best chi square over 10 runs of the algorithm with 100, 200 and 500 ants and 20, 50 and 100 as a tournament size.

However, SNPs in similar regions on the genome are often correlated through a phenomenon known as linkage disequilibrium (LD), and these two SNPs have very similar locations on the genome (36651528 and 36603003) meaning this association is therefore likely to be an artefact of LD, despite it being the strongest statistical signal discovered within this dataset.

There is evidence that this run of the algorithm has found a combination of SNPs that are statistically significantly associated with Type II Diabetes, although LD meant that the result had to be rejected.

There is some evidence that at least one of the SNPs has previously been found to be significant, although not in Type II Diabetes. In [18], the SNP of rs number 7031174 has been found:

The locus rs7031174 resides at the intron of gene MELK. MELK has been found significantly over-expressed in the great majority of breast cancer cells (Involvement of maternal embryonic leucine zipper kinase (MELK) in mammary carcinogenesis through interaction with Bcl-G, a pro-apoptotic member of the Bcl-2 family).

Also the SNPs rs7031174 and rs7045471 are discussed in a previous study [15].

For the combination shown in Table 4, an individual is positive if and only if the SNP rs7031174 is heterozygous and the SNP rs7045471 is common homozygous:

Positive \iff rs7031174 = CA AND rs7045471 = AA

4.2 Combination of SNPs

Although the previous combination is likely to be due to LD, the population-based nature of the ACO algorithm permits the exploration of other combinations that are more promising from a biological perspective. The SNPs that appear individually in tables 5 and 6 and in combination in table 7, are located on different chromosomes and therefore cannot be in LD and demonstrates one of the best results discovered by the algorithm across two chromosomes.

This combination is clearly not as statistically significant as the previous one with a p-value of 4×10^{-15} . However, it is still far in exceedance of the p-values obtained by monte-carlo search (10^{-13}). Additionally, there is evidence in the biology literature that at least one of the SNPs has been previously associated with Type II diabetes. The SNP of rs number 7901695 is in the gene TCF7L2 that is described as follows in [20]:

TCF7L2 transcription factor 7-like 2 (T-cell specific, HMG-box) [Homo sapiens], This gene encodes a high mobility group (HMG) box-containing transcription factor that plays a key role in the Wnt signaling pathway. The protein has been implicated in blood glucose homeostasis. Genetic variants of this gene are associated with increased risk of type II diabetes. Several transcript variants encoding multiple different isoforms have been found for this gene.

The second SNP has not yet been identified as significant

Rs number	position	Chromosome			Allelic OR
7901695	114744078	10			1.371
		TT	TC	CC	
	Diabetics	721	976	296	
	Controls	1367	1333	295	

Table 5: SNP of rs number 7901695.

Rs number	position	Chromosome			Allelic OR
9508846	30306076	13			1.186
		GG	GA	AA	
	Diabetics	1016	722	75	
	Controls	1823	979	128	

Table 6: SNP of rs number 9508846.

	Cases	Controls
Positive	349	843
Negative	1565	2158
Expected Positive	464.2	727.8
Expected Negative	1450	2273
OR	0.571	
95% CI	0.5 to 0.7	
χ^2	61.8	
P-value	4×10^{-15}	

Table 7: Result of the association rs9508846 = Com. Hom. and rs7901695 = Com. Hom..

Combination	p-value
rs4506565=TT AND rs17489797=TT	1.2×10^{-17}
rs12243326=AA AND rs17489797=TT	1.1×10^{-15}
rs9508846=GG AND rs4506565=TT	2.5×10^{-15}
rs9508846=GG AND rs7901695=TT	3.8×10^{-15}
rs4132670=GG AND rs9508846=GG	1.6×10^{-15}
rs7077039=CC AND rs1935683=TT	4.4×10^{-15}
rs4506565=TT AND rs1814811=GG	5.6×10^{-15}
rs7901695=TT AND rs903228=CC	1.4×10^{-14}

Table 8: Sample of the best combinations and their p-value that were discovered.

in diabetes and it does not correspond to any known gene function at present.

For this combination, an individual is positive if and only if the SNP rs9508846 is common homozygous and rs7901695 is common homozygous:

Positive \iff rs9508846 = GG AND rs7901695 = TT

4.3 Sample of the Best Combinations

Table 8 shows a collection of the best results discovered by the ant colony approach on this dataset, again all with significantly smaller p-values than those discovered by the monte-carlo search. Rs4506565 that appears multiple times and rs12243326 are both polymorphisms related to the TCF7L2 gene (in addition to rs7901695 reported above) and therefore have been shown to be related to Type II diabetes. This clearly shows that the ACO algorithm is capable of discovering highly discriminatory associations SNP combinations with some biological plausibility.

5. CONCLUSION

An ant colony approach to the problem of discovering combinations of SNPs from large-scale GWAS data has been

described. The algorithm has been adapted so as to be scalable to the size of dataset both in terms of its memory requirements through the use of a byte-wise representation of genomes and through the use of a tournament path selection to greatly increase execution speeds. These novel modifications to the algorithm are required to permit the execution of ACO on full-scale GWAS data and this is, to the best of our knowledge, the first time that an ACO method has been successfully applied to such data.

The algorithm is also able to discover higher order combinations of SNPs (e.g. 3+ SNPs, not shown) that would not be possible using existing methods. In this paper, combinations of two SNPs that can discriminate type II diabetes patients from controls have been discovered by the approach. The ACO has been able to find some of the strongest statistical signals in the dataset (although as explained above these have been ruled out on biological grounds) and has also found SNPs that have a known biological relationship to diabetes. Although other discovered SNPs do not at present have a known biological function, it is this discovery of plausible known information and targets for further investigation that make the approach a promising addition to the GWAS toolbox.

6. FURTHER WORK

A further step could be to determine which p-values can be expected at random in order to gain a more in-depth understanding of the quality of the results. These p-values can be obtained by running the ACO algorithm that is described here on shuffled data.

The ACO method can be applied to any GWAS dataset that conforms to the standard OXSTATS format and so further trials are planned on other disease datasets from the WTCCC, including Type I Diabetes and Rheumatoid Arthritis.

7. ACKNOWLEDGEMENTS

The work contained in this paper was supported by an EPSRC First Grant (EP/J007439/1).

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from <http://www.wtccc.org.uk>. Funding for the project was provided by the Wellcome Trust under award 076113.

8. REFERENCES

- [1] J. Christmas, E. Keedwell, TM. Frayling, and JRB. Perry. Ant colony optimisation to identify genetic variant association with type 2 diabetes,. In *Information Sciences.*, volume 181, pages 1609–1622, 2011.
- [2] Wellcome Trust Case Control Consortium. Genome-wide association study of 14000 cases of seven common diseases and 3000 shared controls. *Nature*, 447:661–978, 2007.
- [3] Marco Dorigo and Gianni Di Caro. The ant colony optimization meta-heuristic. In *New Ideas in Optimization*, pages 11–32. McGraw-Hill, 1999.
- [4] Sapin E and Keedwell EC. T-aco - tournament ant colony optimisation for high dimensional problems. in *ECTA 2012 - 4th International Conference on*

- [5] M.N. Weedon et al. Combining information from common type 2 diabetes risk polymorphisms improves disease prediction. *PLoS Medicine* 3, 10:1877–1882, 2006.
- [6] Anna L. Gloyn, Michael N. Weedon, Katharine R. Owen, Martina J. Turner, Bridget A. Knight, and Graham Hitman et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell katp channel subunits kir6.2 (kcnj11) and sur1 (abcc8) confirm that the kcnj11 e23k variant is associated with type 2 diabetes. *Diabetes*, 52:568–572, 2003.
- [7] Casey Greene, Bill White, and Jason Moore. Ant colony optimization for genome-wide genetic analysis. In Marco Dorigo, Mauro Birattari, Christian Blum, Maurice Clerc, Thomas Stutzle, and Alan Winfield, editors, *Ant Colony Optimization and Swarm Intelligence*, volume 5217 of *Lecture Notes in Computer Science*, pages 37–47. Springer Berlin / Heidelberg, 2008.
- [8] Kaprio J, Tuomilehto J, Koskenvuo M, Romanov K, Reunanen A, Eriksson J et al. Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in finland. *Diabetologia*, 35(11):1060–7, 1992.
- [9] G. Leguizamón and Z. Michalewicz. A new version of ant system for subset problems. *Proceedings of the 1999 Congress on Evolutionary Computation(CEC 99)*, 2:1458–1464, 1999.
- [10] Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, and Duren WL et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science*, 316(5829):1341–5, 2007.
- [11] Dorigo M., V. Maniezzo, and A. Coloni. Positive feedback as a search strategy. *Tech. Rep. No. 91-016, Politecnico di Milano, Italy*, 1991.
- [12] Kasuga M. Insulin resistance and pancreatic beta cell failure. *J Clin Invest*, 116:1756–1760, 2006.
- [13] J. H. Moore and W. C. White. Exploiting knowledge in genetic programming for genome-wide genetic analysis. in *Lecture Notes in Computer Science*, T. P. Runarsson, H. G. Beyer, E. Burke, J. J. Merelo-Guervós, L. D. Whitley and X. Yao, Eds. Springer, 4193:969–977, 2006.
- [14] Grant SF, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, and Sainz J et al. Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. *Nat Genet*, 38(3):320–3, 2006.
- [15] Dov Shiffman, Devlin J. James, L. Judy Z., and E. Boerwinkle. Detecting two-locus associations allowing for interactions in genome-wide association studies. *Celera Corporation*, 12, 2009.
- [16] Rich SS. Mapping genes in diabetes. genetic epidemiological perspective. *Diabetes*, 39(11):1315–9, 1990.
- [17] T. Stutzle and M. Dorigo. Aco algorithms for the traveling salesman problem 1999. In *Periaux (eds), Evolutionary Algorithms in Engineering and Computer Science: Recent Advances in Genetic Algorithms, Evolution Strategies, Evolutionary Programming, Genetic Programming and Industrial Applications*. John Wiley & Sons, 1999.
- [18] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Nelson L. S. Tang, and Weichuan Yu. Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics*, 26:2517–2525, 2003.
- [19] A.C. Zecchin, H.R. Maier, A.R. Simpson, M. Leonard, and J.B. Nixon. Ant colony optimization applied to water distribution system design: Comparative study of five algorithms. In *Journal of Water Resources Planning and Management*, Vol. 133, No. 1, January 1., 2007.
- [20] Eleftheria Zeggini, Michael N. Weedon, Cecilia M. Lindgren, Timothy M. Frayling, Katherine S. Elliott, and Hana Lango et al. Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science*, 316:1336–1341, 2007.